Ocean Engineering 311 (2024) 118860

Contents lists available at ScienceDirect

Ocean Engineering

journal homepage: www.elsevier.com/locate/oceaneng

Research paper

Deep reinforcement learning-based non-causal control for wave energy conversion

Hanzhen Wang [a], Vincentius Wijaya [b], Tianyi Zeng [c], Yao Zhang [b,*]

[a] Imperial College London, SW7 2BX, United Kingdom
[b] School of Engineering, University of Southampton, SO16 7QF, United Kingdom
[c] Rolls-Royce UTC in Manufacturing and On-Wing Technology, University of Nottingham, NG8 1BB, United Kingdom

ARTICLE INFO

Keywords:
Wave energy converters
Learning-based control
Wave predictions
Double deep Q network

ABSTRACT

As one of the most promising renewable energy resources, ocean wave energy has not been widely commercialized compared to wind energy and solar energy due to its high Levelized Cost of Electricity (LCoE). It has been long recognized that wave energy converter (WEC) control can increase the capture width ratio and enhance the robustness of the WEC against extreme sea states. However, some rigid-body WECs have high nonlinearities and soft-body WECs such as Dielectric Elastomer Generators (DEGs)/Dielectric Fluid Generators (DFGs) can barely be precisely modeled. To tackle these challenges, this paper aims to propose an optimal control scheme that has less dependence on the dynamical model by introducing deep reinforcement learning into the foundation of a non-causal optimal control strategy. The gain parameters are adjusted adaptively in real time to account for an increasing understanding of this scheme on the WEC behavior and the incoming wave. Furthermore, by systematically contrasting outcomes obtained with various prediction time steps, this investigation aims to pinpoint the most effective prediction strategy for optimizing energy capture efficiency. The robustness of the proposed control against prediction errors and model uncertainties has been verified by using the realistic wave data gathered from the coast of Cornwall, UK.

1. Introduction

As a promising renewable resource, wave energy provides high energy density and continuous power supply (Clément et al., 2002; Drew et al., 2009) and has a great potential of supplying global resources of 146 TWh/yr (Kempener and Neumann, 2014). However, compared to wind and solar energy, such potential has not been fully unrealized due to the high Levelized Cost of Energy (LCoE). Various types of wave energy converters (WECs) have been investigated and developed during the past decades, including point absorbers, overtopping WECs, oscillating water columns, and attenuators (Baños et al., 2011), and Dielectric Elastomer Generators (DEGs)/Dielectric Fluid Generators (DFGs) that are developed recently. It has been long recognized that control plays an important role in maximizing energy output and enhancing efficiency. More importantly, it has been proven that wave prediction can further improve the control performance (Falnes and Kurniawan, 2020). Therefore, this paper investigates a non-causal control strategy, in which the current control action is determined by not only the current feedback but also the future information.

Some of the prediction approaches are based on statistical methods, like the Auto-Regressive (AR) prediction method (Zhang et al., 2019) and the extended Kalman Filters (EKF) (Fusco and Ringwood, 2010). As a novel model that has been used in multiple fields, Neural networks have also been introduced to forecast short-term wave forces (Li et al., 2019, 2018). Other prediction methods rely on the extra sensors that can provide measurements of sea wave elevations at multiple upstream locations with certain distances away from the WEC, such as the deterministic sea wave prediction (DSWP) (Abusedra and Belmont, 2011). Recent studies have proposed a large number of non-causal control methods that aim at maximizing wave power production under actuator constraints. These studies show a promising energy harvesting performance (Hals et al., 2010; Li and Belmont, 2014; Ringwood et al., 2014; Genest and Ringwood, 2016; Zhan and Li, 2019). Control methods like MPC based on hydrodynamic principles for WEC control can offer improved performance than traditional control strategy (Faedo et al., 2017). Another study proposed a fully convex implementation, which trades off the energy absorption, the energy consumed by the actuator, and safe operation (Li and Belmont, 2014). A quadratic programming method has been proposed to provide energy maximization solution (Zhong and Yeung, 2018). There are also some

* Corresponding author.
E-mail addresses: hanzhen.wang23@imperial.ac.uk (H. Wang), vvw1n22@soton.ac.uk (V. Wijaya), tianyi.zeng@nottingham.ac.uk (T. Zeng), yao.zhang@soton.ac.uk (Y. Zhang).

https://doi.org/10.1016/j.oceaneng.2024.118860
Received 18 March 2024; Received in revised form 12 July 2024; Accepted 28 July 2024
Available online 6 August 2024

other effective methods proposed (Ringwood et al., 2014; Korde and Ringwood, 2016; Fusco and Ringwood, 2012). However, these methods rely on the accurate WEC model.

Machine learning techniques have shown amazing performance in conducting complex tasks, especially facing ambiguous inputs like nature language (Collobert and Weston, 2008), image classification (Krizhevsky et al., 2012), and data-driven modeling. Machine learning comes into the WEC control community mainly in two methods. The first is to use machine learning methods to build a data-based, nonlinear model of the system dynamic for system identification (Gaspar et al., 2016). The second is to optimize the paraments (Anderlini et al., 2017) of other control theories or to conduct data-driven online control (Tri et al., 2016). Different from supervisor learning and non-supervisor learning, reinforcement learning lets the agent learn from the interaction with the environment (Kober et al., 2013). Based on data-driven logic, reinforcement learning has shown satisfactory performance in dealing with systems with uncertainty (Sutton and Barto, 2018), like games (Mnih et al., 2013) and go (Silver et al., 2016), etc. This makes the reinforcement learning method suitable to deal with inaccurate predictions and new circumstances that the agent has not met before. Therefore, reinforcement has been used in robotics systems (Kober et al., 2013), which have similar mechanics to the WEC problem and are faced with many uncertain circumstances. A study has shown it has the potential to tackle the inaccuracy of the WEC control problem (Zou et al., 2022). However, the research on the application of reinforcement learning in the WEC field is still insufficient. Only a few studies of WEC involve reinforcement learning in WEC control (Zou et al., 2022; Anderlini et al., 2018; Bruzzone et al., 2020). The Q-learning algorithm is one of the classical value-based RL algorithms (Watkins and Dayan, 1992). The research (Anderlini et al., 2016) by Anderlini et al. applies Q-Learning in identifying the optimal damping for WECs. Due to the complexity of a WEC system, Double Deep Q-Network (DDQN), a kind of Deep Reinforcement Learning (DRL), is introduced to WEC control later (Anderlini et al., 2020). The DDQN algorithm was put forward to play artari (Mnih et al., 2013) at first but shows great competence in improving the performance of the Q-learning algorithm. Usually, a DDQN method is used to solve discrete problems, but a recent study applied a time-varying PD control whose gains are adjusted by DDQN (Zou et al., 2022).

This paper investigates a non-causal control strategy using DDQN developed for point absorber WEC systems, whose control performance is improved by benefitting from both the wave prediction and the DDQN. This paper aims to fill the research gap of incorporating wave prediction into model-free control methods like DDQN. Although the future information of waves could be partly reflected by the prediction ability of DDQN (Zou et al., 2022; Anderlini et al., 2020), this kind of prediction only works to refer to the future reward. DDQN always makes the best decision based on the present reward and the estimate of future reward, but the control of PTO is conducted by the time-variant PD controller. Wave prediction can still improve the performance of the PD controller. Besides, a comparison of the improvement brought by different wave prediction horizons is also necessary to show how prediction influences control performance. The contributions of this paper are as follows

- A non-causal control based on the DDQN is proposed to fill the research gap of the prediction-based reinforcement learning in wave energy conversion.
- Wave prediction is incorporated into the control scheme to increase the energy output. Taking the benchmark problem, a point absorber, as an example, it has been proven that wave prediction makes a significant contribution to wave energy harnessing.
- A pair of real-time adjustable gain parameters have been optimized via reinforcement learning to tackle the challenges in wave energy converters with significant nonlinearities and model mismatch.

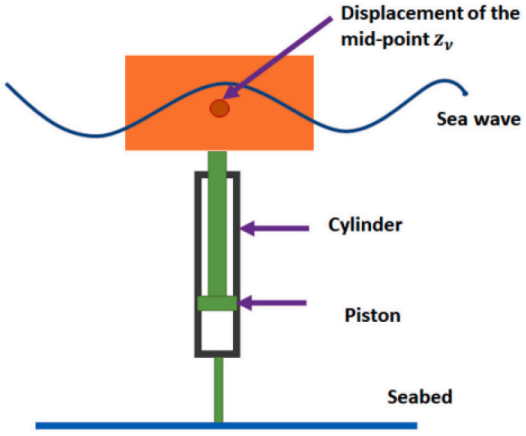

Fig. 1. Schematic diagram of the point absorber.

- A realistic wave data gathered from the coast of Cornwall, Wales, UK is used to validate the effectiveness of the proposed control algorithm.
- The proposed control algorithm is generally applicable to other WECs across varied archetypes (e.g., sizes, shapes), especially for flexible WECs, such as origami WECs and Dielectric Elastomer Generator and Dielectric Fluid Generator.
- The proposed method is robust against model uncertainties and prediction errors.

The rest of the paper is as follows. Section 2 introduces the state–space model of the point absorber. The reinforcement learning control method is proposed in Section 3, where the basic structure of the agent is introduced. Simulation results for the comparison are shown in Section 4. Section 5 concludes this paper.

2. WEC modeling for environment training

This section first introduces the dynamical model of a single-point absorber in Section 2.1. To build the simulation environment, the hydrodynamic model is described in a state–space model form, which introduces modeling uncertainties. The Section 2.2 shows how the hydrodynamic model is transformed into a state–space model that is used in training the environment. The Section 2.3 gives the optimal solution of a non-causal WEC control problem for a precisely modeled point absorber.

2.1. Dynamic model

Fig. 1 shows part of a possible hydraulic power take-off (PTO) design where a hydraulic cylinder is vertically installed below the float and is fixed to the bottom of the seabed. More details on this design can be found in Weiss et al. (2012). z_w and z_v are the water level and the height of the mid-point of the float respectively. The PTO torque is proportional to the force f_u acting on the piston inside the cylinder. The extracted power is $P = -f_u v$, where the velocity on the piston is $v = \dot{z}_v$.

By using Newton's second law, the dynamic equation (Yu and Falnes, 1995) for the float of the point, the absorber is like function (1).

$$m_s \ddot{z}_v = -f_s - f_r + f_e + f_u \tag{1}$$

where m_s is the float mass. The restoring force f_s is given by Eq. (2).

$$f_s = k_s z_v \tag{2}$$

With the hydrostatic stiffness $k_s = \rho g_s$, ρ as water density, g as standard gravity, and s as the cross-sectional area of the float. f_r is the radiation force determined by Eq. (3).

$$f_r = m_\infty \ddot{z}_v + \int_{-\infty}^{\infty} h_r(\tau)\dot{z}_v(t-\tau)\,d\tau \tag{3}$$

where m_∞ is the added mass, h_r is the kernel of the radiation force that can be computed via hydraulic software packages (e.g. WAMIT (Lee, 1995)). Following (Yu and Falnes, 1995), the convolutional term in (3) can be approximated by a causal finite-dimensional state–space model.

$$\begin{aligned} \dot{x}_r &= A_r x_r + B_r \dot{z}_v \\ f_R &= C_r x_r \approx \int_{-\infty}^{t} h_r(\tau)\dot{z}_v(t-\tau)\,d\tau \end{aligned} \tag{4}$$

where $(A_r, B_r, C_r, 0)$ and $x_r \in \mathbb{R}^{n_r}$ are the state–space realization and the state respectively. Following (Yu and Falnes, 1995), the wave excitation force f_e can be determined by (5).

$$f_e = \int_{-\infty}^{\infty} h_e(\tau) z_w(t-\tau)\,d\tau \tag{5}$$

where h_e is the kernel of the radiation force and the state–space approximation is given by

$$\begin{aligned} \dot{x}_e &= A_e x_e + B_e z_w \\ f_e &= C_e x_e \approx \int_{-\infty}^{t} h_e(\tau) z_w(t-\tau)\,d\tau \end{aligned} \tag{6}$$

where $(A_e, B_e, C_e, 0)$ and $x_e \in \mathbb{R}^{n_e}$ are the state–space realization and the state respectively.

2.2. State–space model

With the realizations of (4) and (6), the state–space model of (1) can be represented by

$$\begin{cases} \dot{x} &= A_c x + B_{uc} u + B_{wc} w + \epsilon \\ y &= C_c x \end{cases} \tag{7}$$

where $w = z_w$ is the wave elevation whose prediction is incorporated into the controller design, $y = z_v$, $y = \dot{z}_v$, $x = [z_v, \dot{z}_v, x_r, x_e]$, $u = f_u$. ϵ represents the modeling uncertainty caused by wave force approximations (4) and (6). And

$$\begin{aligned} A_c &= \begin{bmatrix} 0 & 1 & 0 & 0 \\ -\frac{k_s}{m} & 0 & \frac{C_e}{m} & -\frac{C_f}{m} \\ 0 & B_r & A_r & 0 \\ 0 & 0 & 0 & A_e \end{bmatrix} B_{wc} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ B_e \end{bmatrix} B_{uc} = \begin{bmatrix} 0 \\ \frac{1}{m} \\ 0 \\ 0 \end{bmatrix} \\ C_c &= \begin{bmatrix} 0 & 1 & 0_{1\times(n_r+n_e)} \end{bmatrix} \end{aligned} \tag{8}$$

with $m = m_s + m_\infty$. The continuous-time model (7) can be converted to a discrete-time model (9).

$$\begin{cases} x_{k+1} &= A x_k + B_u u_k + B_\omega \omega_k + \epsilon_k \\ y_k &= C x_k \end{cases} \tag{9}$$

where the pair (A, B_u, B_ω, C) is the discrete-time form of the pair $(A_c, B_{uc}, B_{\omega c}, C_c)$.

2.3. Linear Optimal control for a precisely modeled point absorber

In the Linear Optimal control of the point absorber (Zhan and Li, 2019), the following constraints in Eqs. (10a) and (10b) should be satisfied.

$$|z_v| \le \Phi_{max} \tag{10a}$$

$$|f_u| \le u_{max} \tag{10b}$$

The optimal control strategy of a WEC based on point absorber can be formulated as:

$$\min \frac{1}{N}\left(\sum_{k=0}^{N-1} L_k(x_k, u_k)\right) \tag{11a}$$

$$s.t. x_{k+1} = A x_x + B_u u_k + B_\omega \omega_K \tag{11b}$$

$$z_k = C_z x_k \tag{11c}$$

where N is the number of prediction steps and L_k is the stage cost. And we have,

$$L_k = \frac{1}{2} {x_k}^T Q x_k + z_k u_k + \frac{1}{2} r {u_k}^2 \tag{12}$$

The following criteria hold.

1. $\frac{1}{2}{x_k}^T Q x_k$ is used to penalize the state. The weight Q influences the stability of the control system and can be used as a tuning parameter to handle the state constraint (10a).
2. $-z_k u_k$ represents the power that can be captured by the PTO mechanism.
3. $\frac{1}{2} r {u_k}^2$ is used to penalize the input. The weight r influences the stability of the control system and can be used as a tuning parameter to handle the input constraint (10b).

The linear optimal noncausal controller for this control problem can further be simplified by linear optimal noncausal control law (13), according to Zhan and Li (2019).

$$u_k = K_p^* x + K_d^* \omega \tag{13}$$

where ω is a vector of wave elevation prediction, and gain parameters K_p^* and K_d^* are calculated offline.

For a precisely modeled point absorber, the linear non-causal control (13) has been proven to be effective and efficient. However, for WECs with significant model mismatch and nonlinearities in hydrodynamics, such as DEGs and DFGs as well as origami WECs, the gain parameters should be adjusted in real-time to deal with the challenges in model-less or model-free optimization problems. In this paper, a non-causal control scheme based on a reinforcement learning agent is proposed to tackle the challenge.

3. Non-causal control with reinforcement learning

3.1. Deep reinforcement learning control framework

Reinforcement learning could be generally divided into model-based reinforcement learning and model-free reinforcement learning. Unlike traditional MPC, model-free reinforcement control does not require a precise model for controller design. Instead, a reinforcement learning control system uses the interaction with the environment to learn the policy. Reinforcement learning has shown a strong ability to deal with sequential decision-making (Li, 2017), where the problem of the non-causal control for WEC falls in. The DDQN (Mnih et al., 2013) is now widely used in deep reinforcement learning. It is developed from classical Q-Learning which has been proven to be effective in simple problems. Based on Q-Learning DDQN introduced deep neural networks (DNN) to instead the Q-table used in traditional Q-Learning. This leads the RL algorithm to gain the ability to solve more complex problems like chess and video games.

In a typical reinforcement learning (RL) system (Sutton and Barto, 1998), there are an "agent" and an "environment". Due to the high cost of the mechanical structure, simulation environments are always applied to train the agent. For a DDQN agent, there is a DNN to be trained. In an RL problem, we use s_n to present the current state. Meanwhile a_n stands for the current action the agent does to the environment. The action is selected in an action space according to the policy. Besides, r_n is the reward defined manually to critique the performance of the current step. The calculation of reward always requires professional knowledge. After one step of interaction, the system goes forward and we get s_{n+1}, a_{n+1} and r_{n+1}. In each step, the selection of action is regarded as a Markov decision process. The decision is based on the value function. There are two neural networks with wight parameters

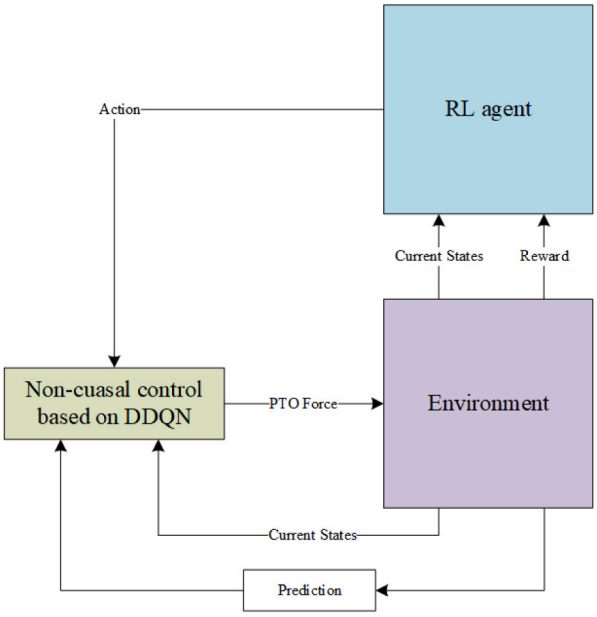

Fig. 2. Structure of the control system.

θ and θ^-, which have totally the same structure Q. Thus, the target y_T is expressed as (14).

$$y_T = r_n + \gamma \max_{a_{n+1}} Q(s_{n+1}, a_{n+1}; \theta_n^-) \tag{14}$$

where γ is the future reward discount that represents how much we focus on the future reward. The agent is expected to learn the policy to maximize the target y. The weight of Q-network(θ) will be updated based on the weight of the target network (θ^-) with Eq. (15).

$$\nabla_{\theta_n} L(\theta_n) = E[(y_T - Q(s_n, a_n, \theta_n))\nabla_{\theta_n} Q(s_n, a_n, \theta_n)] \tag{15}$$

Furthermore, a minibatch training (sampled from the stored experience buffer) is also adopted to avoid divergence and smooth the learning. The agent collects r_n and new state s_{n+1} after the last action a_n has been taken. The experience is set $e_n = [s_n, a_n, r_n, s_{n+1}]$. The experience of the agent is saved in a buffer to improve the learning speed. The batch is sampled from the experience buffer to train the deep network with (15). Then the next action a_{n+1} is decided by maximizing the target with (14). The target network θ_n^- is updated after each batch by the Q-network with the function (16).

$$\theta_n^- = \tau\theta_n + (1-\tau)\theta_n^- \tag{16}$$

where τ is the smoothing factor.

3.2. DDQN problem formulation

Considering the importance of wave prediction in WEC control, the WEC control is formulated in the fashion of time-varying PD control with prediction, which matches the form of the optimal solution proposed by Zhan and Li (2019). Fig. 2 shows the basic structure of the system. The TVPD controller can be implemented by (17).

$$F_{PTO} = K_p(n)s_n + K_d(n)w_{pre} \tag{17}$$

where K_p and K_d are adjusted in real-time by the action of RL, and the w_{pre} is the wave prediction. s_n contains the displacement of float $z_{v,n}$ and the velocity of the float $\dot{z}_{v,n}$ at the step n. The state can be expressed as Eq. (18).

$$s_n = [z_{v,n}, \dot{z}_{v,n}]^T \tag{18}$$

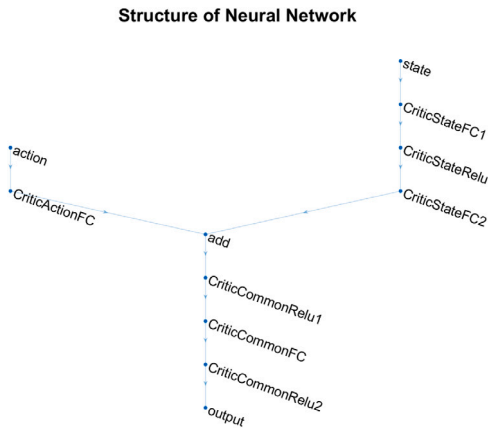

Fig. 3. Structure of the DNN.

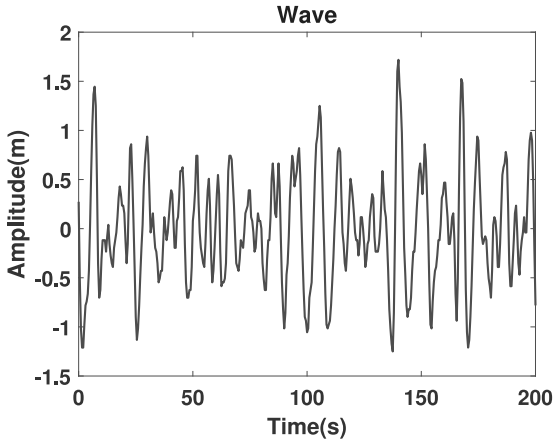

Fig. 4. Wave height in the first 200 s.

And the length of w_{pre} is the prediction horizon N.

$$w_{pre} = [w_1, w_2, \ldots, w_N]^T \tag{19}$$

with N as the prediction horizon, which is a positive integer.

To balance the cost of computing and control performance, the control period of the RL agent and the sampling period are set at the same $T_s = 0.5s$. When the RL agent does not work between two sample points, the PD controller can still work as a traditional PD control. Therefore, the mission of the RL agent is to adjust K_x and K_d every RL sample period. The action of the RL algorithm is from the given action space A.

$$Action_n = \{a_n|[(\delta, 0, \ldots, 0), (-\delta, 0, \ldots, 0), (0, \delta, \ldots, 0), (0, -\delta, \ldots, 0), \ldots]\} \tag{20}$$

where δ is a small amount. The length of each action choice is decided by the length of the control horizon. The given action space is similar to one-hot code to some degree. In each step, the non-causal control is updated by the rule (21)

$$[K_p(n+1), K_d(n+1)] = [K_p(n), K_d(n)] + a_n \tag{21}$$

According to a previous study (Anderlini et al., 2018), there is no need to consider more complex actions like $(\delta, \delta, \ldots)$ and $(-\delta, -\delta, \ldots)$ to concise the action space. In a WEC control system, we focus on the energy output efficiency most. So it is reasonable to set a reward related to the power P_n. Whilst we need to protect the mechanical structure so we set a "punishment" r_{punish} when the F_{PTO} output exceed the

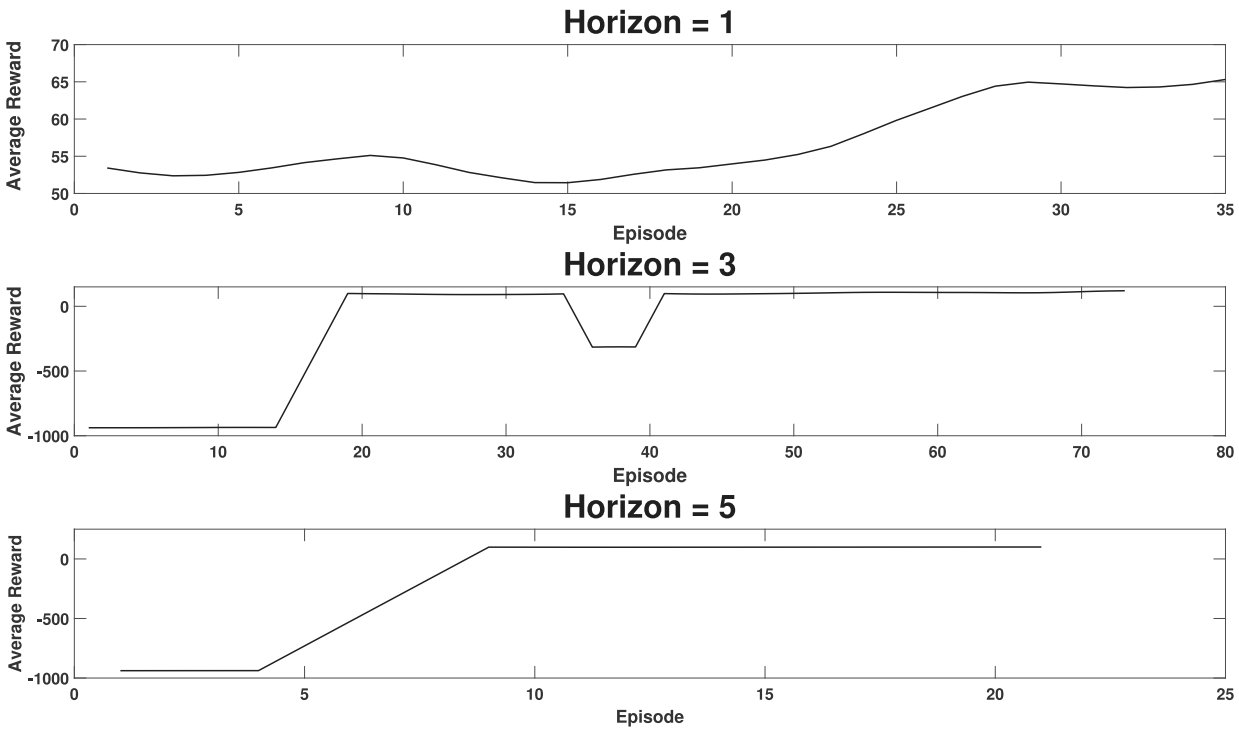

Fig. 5. Average reward during training.

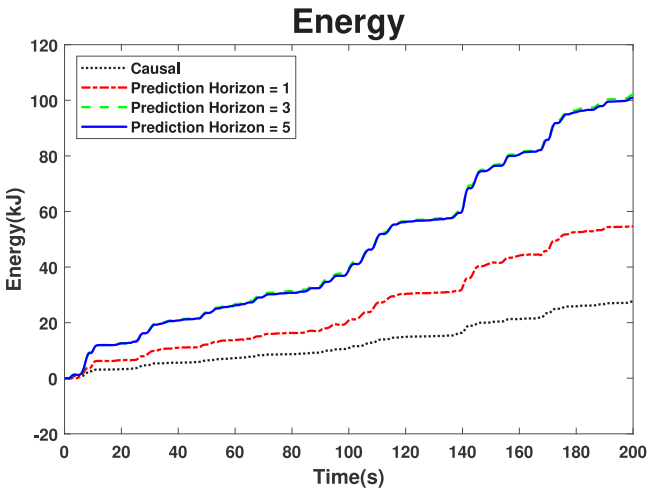

Fig. 6. Energy output by different horizons.

limitation $F_{\max}$. Therefore, the reward r_n is expressed like (22).

$$r_n = \begin{cases} P_n, & |F_{PTO}| \le F_{max} \\ r_{punish}, & |F_{PTO}| \ge F_{max} \end{cases} \tag{22}$$

In this case, r_{punish} is always set to a very small plural number to let the agent learn to avoid the behavior that will damage the system. Also, it is worth mentioning that the period of RL sampling and the period of simulation is different. Thus, the P_n is actually calculated as Eq. (23).

$$P_n = \frac{P\delta t}{\delta t_{RL}} \tag{23}$$

Through the Q-function, the DRL control method can take all the future discount rewards which consist of past rewards and future rewards. This is similar to $\sum r_n$ but uses an estimate of the future.

Table 1
Training parameters.

Hyperparameters	Values
Learning Rate	0.01
Batch Size	64
Replay Memory Size	10^4
Discount Factor	0.999
Target smoothing Factor	0.99
Number of Neurons Each Hidden Layer	24
Target Update period	0.5s

3.3. DQN Agent structure and training

The structure of the neural network is like Fig. 3. Each fully connected layer contains 24 cells and the optimizer is SGD. The agent needs to be trained previously. Parameters are set as Table 1. Besides, we train three sets of agents facing different forecast horizons. For the agent in each set, the initial value of the TVPD controller is set by the linear optimal noncausal control method proposed in Zhan and Li (2019).

3.4. Wave prediction

The effectiveness of wave prediction techniques is crucial for the proposed noncausal optimal WEC control strategy. Wave prediction methods can generally be classified into two main categories.

The first category relies on historical sea wave data collected at the same location as the WEC. These methods utilize statistical approaches such as autoregressive models, cyclical models, and extended Kalman filters, as summarized in Fusco and Ringwood (2010). Also, wave prediction can be obtained by learning-based algorithms such as Long-Short Term Memory LSTM (Zhang et al., 2021) and ARIMA (Wu et al., 2021).

The second category of wave prediction techniques uses measurements of sea wave elevations from multiple nearby locations at certain distances from the WEC. By analyzing wave propagation and direction, these methods can predict the sea wave profile at the WEC's location several seconds into the future. A notable method in this category is

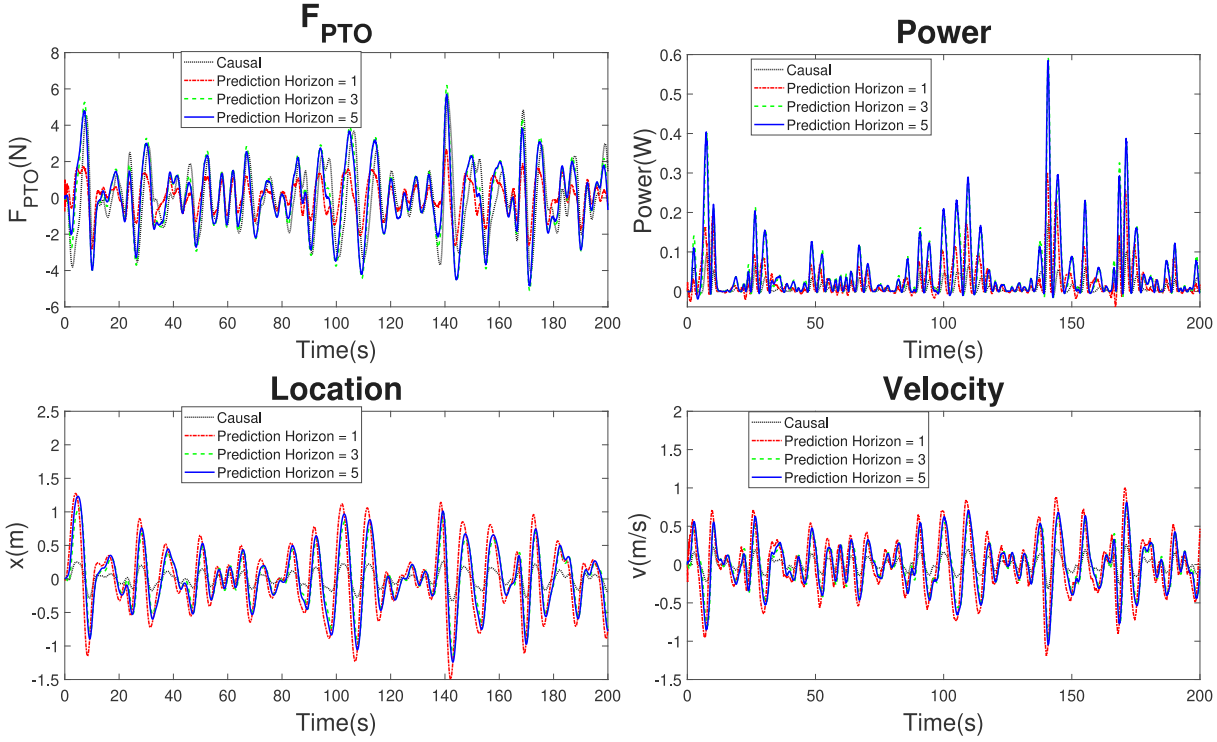

Fig. 7. F_{PTO}, power, location and velocity.

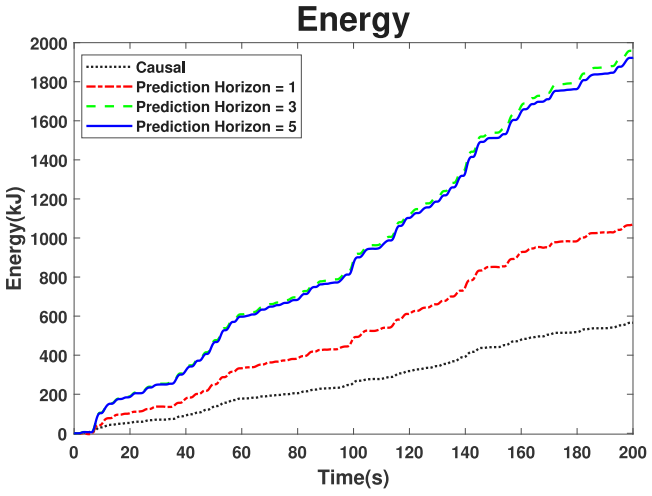

Fig. 8. Energy output on computer-generated data.

Deterministic Sea Wave Prediction (DSWP), which we describe briefly for completeness. More detailed information can be found in Belmont et al. (2014).

4. Simulation results

This section presents the results of the simulation, conducted using MATLAB for both the simulation environment and the agent. The wave data we used is gathered from the coast of Cornwall, Wales. The wave height in the first 200 s is shown in Fig. 4.

Three sets of simulations were established in this section, each with prediction horizons set at 1, 3, and 5 steps. Each simulation episode comprised 20,001 steps, equivalent to 200 s in a realistic scenario. The agent's training ceased when the total rewards reached a relatively stable value. The three simulation sets required 35, 73, and 21 episodes, respectively. The training duration for the agent was approximately 6 h, 12 h, and 5 h. The hardware employed for training included an Nvidia GTX1650 GPU and an Intel Core i7-9750H CPU. The average reward, filtered with a moving average filter for enhanced clarity, is depicted in Fig. 5. This visualization provides insights into the learning progress of the agent across different prediction horizons and demonstrates the efficiency of the training process.

Table 2
Energy output of WEC with different controllers.

Controller	Energy output
DDQN with prediction horizon 5	100.9 kJ
DDQN with prediction horizon 3	102.0 kJ
DDQN with prediction horizon 1	54.7 kJ
Traditional PD controller	27.64 kJ

In the context of a WEC system, the primary focus lies on the converted energy. As a benchmark for comparison, we utilize the conventional control method to highlight the performance differences. The total energy output after 200 s simulation is shown in the Table 2:

The energy output from both the classical control methods and our learning-based control method is presented in Fig. 6. Additionally, we conduct a comparative analysis of outcomes achieved through different prediction horizons, providing insights into the impact of varying horizons on energy conversion.

As illustrated in Fig. 6, there is a remarkably similar performance between Horizon 3 and Horizon 5. Notably, it is worth highlighting that the training cost for the agent with Horizon 5 is lower, as depicted in Fig. 5. This underscores the efficiency of our method, surpassing traditional control methods. In reality, WEC systems operate under various constraints. Beyond energy output, additional criteria such as F_{PTO}, velocity, and location are crucial considerations, as shown in Fig. 7. The comprehensive evaluation of these factors reaffirms the effectiveness of our approach, demonstrating its superiority over traditional control methods and its potential for practical implementation in real-world WEC systems.

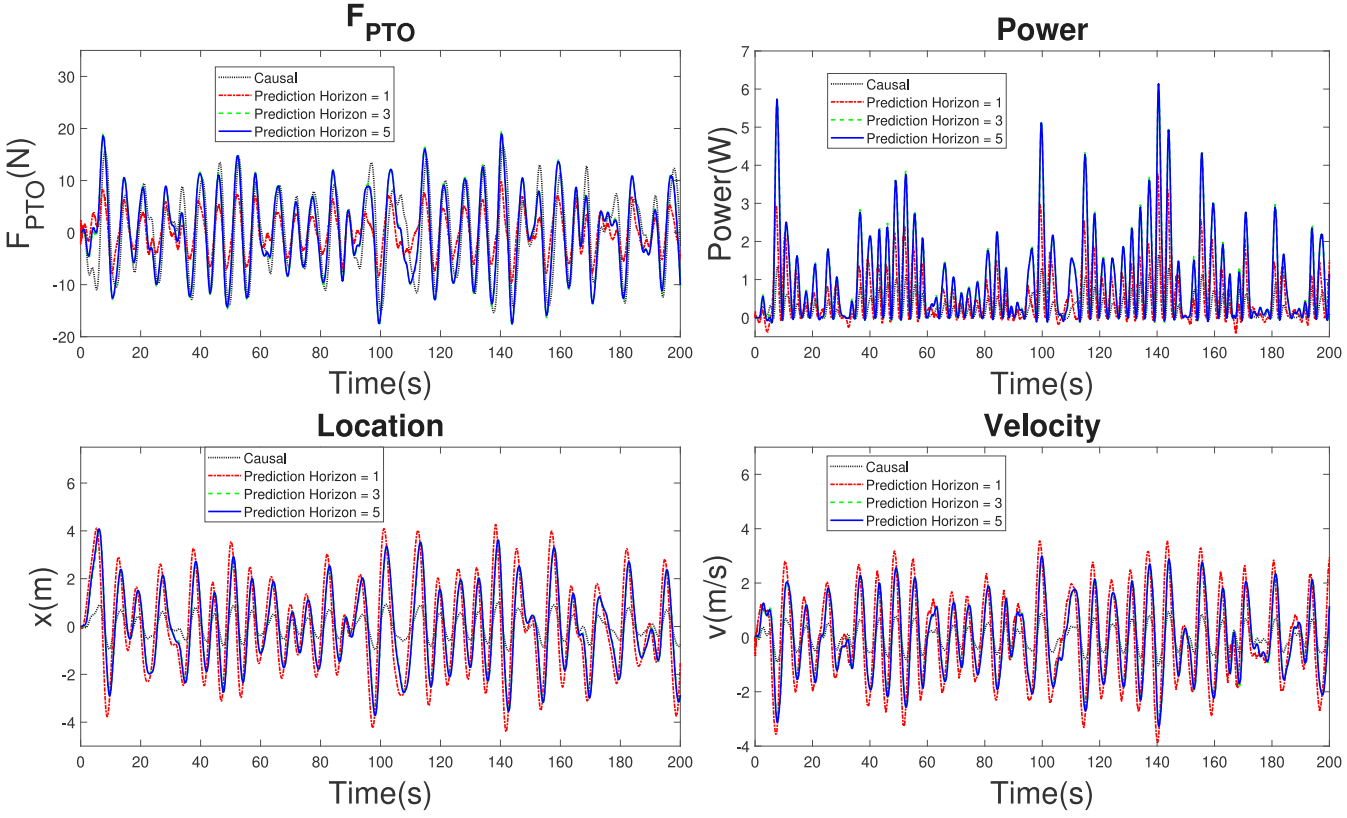

Fig. 9. F_{PTO}, power, location and velocity.

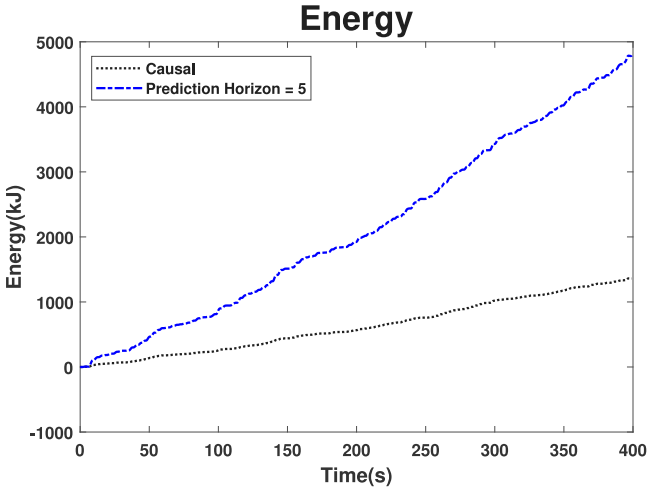

Fig. 10. Simulation in 400 s.

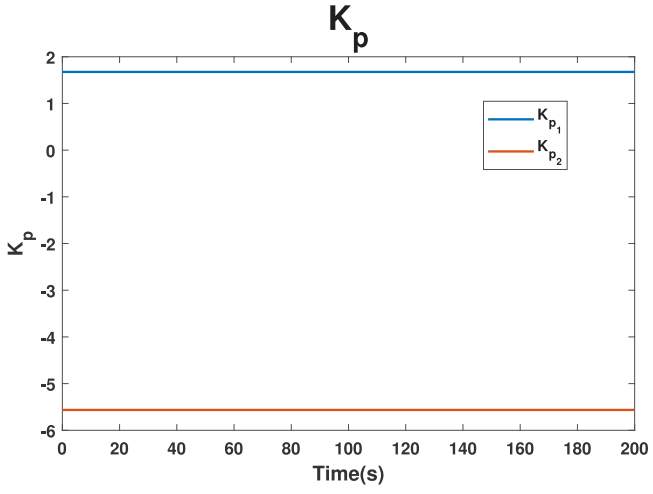

Fig. 11. The gain K_p vs time (K_{p1} is the gain for the displacement of the float, and K_{p2} is the gain for the velocity of the float).

Analyzing Fig. 7, it is evident that all the factors, including F_{PTO}, power, location, and velocity, are effectively controlled. Notably, an interesting trend emerges from this data with an increase in the prediction step size, there is a corresponding increase in the F_{PTO} of the system. Simultaneously, the displacement and speed of the float exhibit a decrease as the prediction horizons expand. This observation implies that larger prediction step sizes contribute to greater energy output. The interconnected relationship between these variables suggests that adjusting the prediction horizons can influence the system dynamics, striking a balance between maximizing energy extraction and controlling the mechanical forces within acceptable limits. This insight is crucial for fine-tuning the agent's parameters to achieve optimal performance under varying conditions.

To assess the universality of the trained agent, acquiring additional data becomes imperative. However, obtaining real sea wave data is not always convenient. To address this challenge, computer-generated data can facilitate more extensive testing. Furthermore, considering the

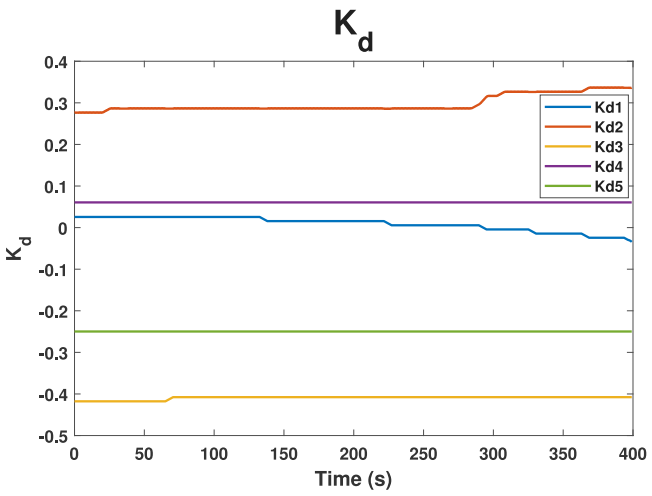

Fig. 12. The gain K_d vs time (K_{di} is the gain for i th-step-ahead prediction).

variability in wave amplitudes under different sea conditions, training the agents under more challenging conditions than the simulation becomes crucial. This approach ensures that the force exerted by the F_{PTO} remains within mechanical limitations. Under the evaluation using generated wave data, the power output is illustrated in Fig. 8, while Fig. 9 provides insights into the values of position, velocity, power, and F_{PTO}. This comprehensive testing approach, incorporating simulated conditions and tougher training scenarios enhance the agent's adaptability to a broader range of situations, contributing to its overall universality.

The wave is intentionally set to a more violent setting to assess the agent's performance under harsher environmental conditions. As depicted in the figure, the energy output increases proportionally with the heightened energy content of the wave. Notably, the constraint on the value of F_{PTO} remains strict, capped at $20N$. Additionally, extending the prediction horizon leads to improved energy output, demonstrating that a longer prediction duration enhances the algorithm's performance. This effect is particularly evident when considering the stability of the algorithm over time. As illustrated in Fig. 10, the total energy output significantly surpasses that of the conventional method when the time window is doubled. This observation underscores the algorithm's ability to adapt to varying wave intensities and highlights the positive correlation between prediction horizon length and energy harvesting performance. The strict adherence to the $20N$ constraint ensures the agent's operation within defined safety limits, emphasizing both efficiency and safety in energy extraction.

During the operation of the system, both K_p and K_d undergoes continuous changes. The dynamic evolution of K_p and K_d gain is illustrated in Figs. 11 and 12, providing an overview of its general trend over time.

In this particular scenario, the simulation results indicate that the parameter K_{d_2} undergoes the most rapid changes among the specified parameters. While K_p keeps constant and other parameters have very small changes. This observation suggests that the second step in the future holds significant importance, potentially influencing the system dynamics to a considerable extent. Moreover, the parameter K_{d_1} exhibits substantial variations over time, implying that the immediate future carries a higher level of reliability and impact on the system. This observation may shed light on the comparable performance of the horizon 3 and horizon 5 cases in terms of energy output. The dynamic nature of K_{d_2} and the pronounced changes in K_{d_1} over time might contribute to the convergence of performance between these two horizon scenarios. This nuanced relationship is further illustrated in Fig. 13, highlighting the correlation between K_{d_2} and the running time. It is noteworthy that both Figs. 12 and 13 present data processed through the move-mean technique, ensuring a smoothed representation of the underlying trends and patterns in the simulation results. This analytical approach enhances the clarity and interpretability of the depicted data.

Table 3
Energy output.

(horizon = 5)	K = 0	K = 0.05	K = 0.10
SNR = 0 dB	100.2 kJ	105.1 kJ	115.1 kJ
SNR = 10 dB	100.3 kJ	107.5 kJ	126.6 kJ
SNR = 20 dB	100.2 kJ	108.2 kJ	124.1 kJ

In practice, model uncertainty and prediction error are inevitable. To simulate prediction error, we introduce white noise to the prediction signal. To account for model uncertainty, we employ a model that incorporates this uncertainty in Eq. (9). To test the performance of our agent, we change the test environment without retraining the agent. By adding noise with a signal-to-noise ratio (SNR) equal to 0 dB, 10 dB and 20 dB to the horizon 5 case, we get Fig. 14.

From Fig. 14, we can find the noise added has a very slight influence. This might result from the build-in robustness of optimal control and the controller does not highly rely on the prediction or the influence of the five steps offset each other. Then by adding a random uncertainty $\epsilon_k = K\epsilon(k)$ to the model, we get the result shown in Fig. 15. Where K is a coefficient and $\epsilon(k)$ is a random variable evenly distribute between -1 and 1.

Fig. 15 shows that with model uncertainty, the system can still keep stable. Table 3 shows the final energy output of different combinations of uncertainty and prediction error.

We can find from the above table that the proposed method is effective and robust against model uncertainties and prediction errors under the condition that there are acceptable noise below 20 dB and model uncertainty under 0.1 (m/s).

5. Conclusions

In light of the unique characteristics of Wave Energy Converters (WECs), we have introduced a reinforcement learning methodology to augment the non-causal controller based on Double Deep Q Network (DDQN). After subjecting our approach to rigorous simulation tests using the collected dataset, we have achieved energy conversion results that outperform traditional control methods. Furthermore, we conducted a comparative analysis of the control effects stemming from different prediction time steps.

Our findings unequivocally demonstrate that longer-term predictions, specifically, those involving three steps at 0.03 s and five steps at 0.05 s, significantly outperform the results obtained from shorter-term predictions, such as one step at 0.01 s. This indicates that extending the prediction horizon has a profound impact on improving the controller's efficacy. Additionally, it is noteworthy that even though a five-step prediction has a negligible effect on the ultimate total energy output, it does alleviate the training burden on the reinforcement learning agent. This observation suggests that the increased predictive horizon enhances the controller's performance, consequently lightening the workload for the reinforcement learning agent.

In conclusion, by integrating a reinforcement learning approach and extending the predictive horizon, we have achieved remarkable progress in WEC control strategies. These advancements hold the promise of significantly enhancing energy conversion efficiency and contributing to the sustainability of energy production.

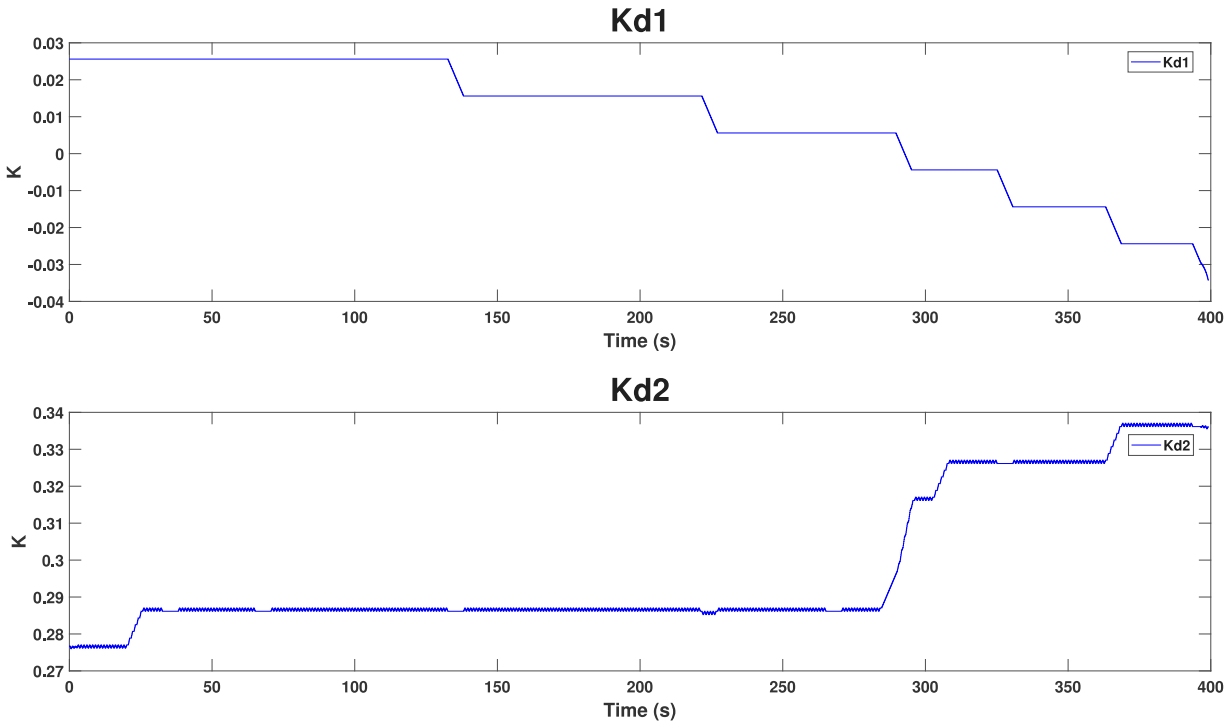

Fig. 13. K_{d_1} and K_{d_2}.

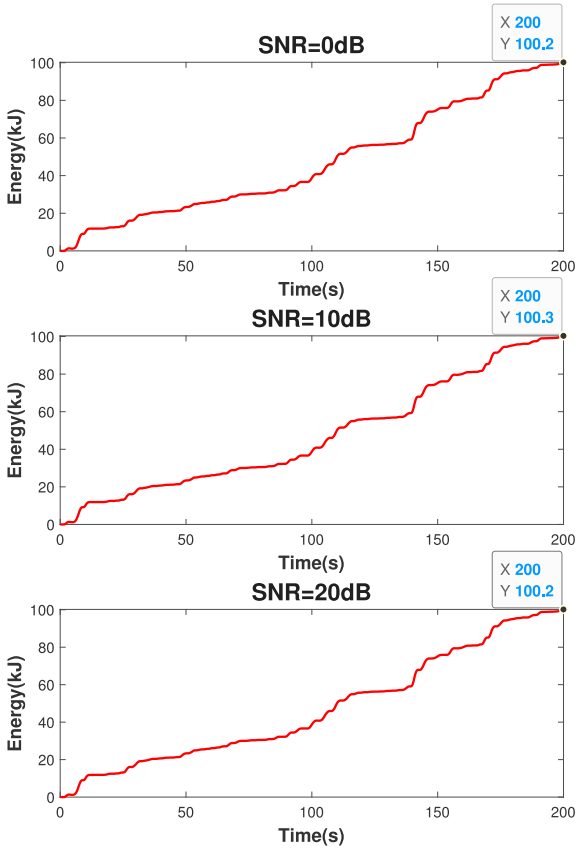

Fig. 14. Energy output with SNR equal 0 dB, 10 dB, 20 dB.

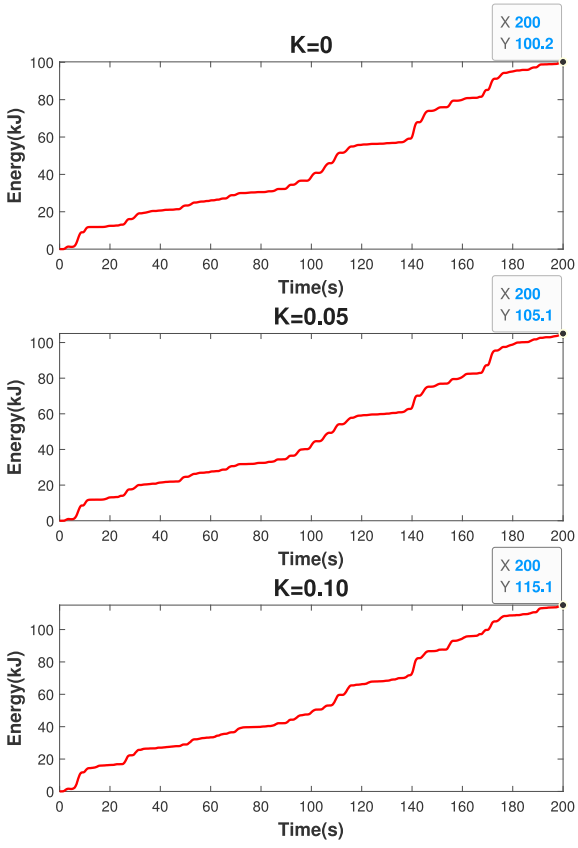

Fig. 15. Energy output with $K = 0, 0.05, 0.1$.

CRediT authorship contribution statement

Hanzhen Wang: Methodology, Conceptualization. **Vincentius Wijaya:** Conceptualization. **Tianyi Zeng:** Validation, Formal analysis. **Yao Zhang:** Supervision, Resources, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was funded by Wave Energy Scotland Direct Generation Competition and the UK Royal Society IEC-NSFC (223485).

References

Abusedra, L., Belmont, M., 2011. Prediction diagrams for deterministic sea wave prediction and the introduction of the data extension prediction method. Int. Shipbuild. Prog. 58 (1), 59–81.

Anderlini, E., Forehand, D., Bannon, E., Abusara, M., 2017. Reactive control of a wave energy converter using artificial neural networks. Int. J. Mar. Energy 19, 207–220.

Anderlini, E., Forehand, D., Bannon, E., Xiao, Q., Abusara, M., 2018. Reactive control of a two-body point absorber using reinforcement learning. Ocean Eng. 148, 650–658. http://dx.doi.org/10.1016/j.oceaneng.2017.08.017.

Anderlini, E., Forehand, D.I.M., Stansell, P., Xiao, Q., Abusara, M., 2016. Control of a point absorber using reinforcement learning. IEEE Trans. Sustain. Energy 7 (4), 1681–1690. http://dx.doi.org/10.1109/TSTE.2016.2568754.

Anderlini, E., Husain, S., Parker, G.G., Abusara, M., Thomas, G., 2020. Towards real-time reinforcement learning control of a wave energy converter. J. Mar. Sci. Eng. 8 (11), http://dx.doi.org/10.3390/jmse8110845.

Baños, R., Manzano-Agugliaro, F., Montoya, F., Gil, C., Alcayde, A., Gómez, J., 2011. Optimization methods applied to renewable and sustainable energy: A review. Renew. Sustain. Energy Rev. 15 (4), 1753–1766. http://dx.doi.org/10.1016/j.rser.2010.12.008.

Belmont, M., Christmas, J., Dannenberg, J., Hilmer, T., Duncan, J., Duncan, J., Ferrier, B., 2014. An examination of the feasibility of linear deterministic sea wave prediction in multidirectional seas using wave profiling radar: Theory, simulation, and sea trials. J. Atmos. Ocean. Technol. 31 (7), 1601–1614.

Bruzzone, L., Fanghella, P., Berselli, G., 2020. Reinforcement learning control of an onshore oscillating arm Wave Energy Converter. Ocean Eng. 206, 107346. http://dx.doi.org/10.1016/j.oceaneng.2020.107346.

Clément, A., McCullen, P., Falcão, A., Fiorentino, A., Gardner, F., Hammarlund, K., Lemonis, G., Lewis, T., Nielsen, K., Petroncini, S., Pontes, M.-T., Schild, P., Sjöström, B.-O., Sørensen, H.C., Thorpe, T., 2002. Wave energy in Europe: current status and perspectives. Renew. Sustain. Energy Rev. 6 (5), 405–431. http://dx.doi.org/10.1016/S1364-0321(02)00009-6.

Collobert, R., Weston, J., 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine Learning. ICML '08, Association for Computing Machinery, New York, NY, USA, pp. 160–167. http://dx.doi.org/10.1145/1390156.1390177.

Drew, B., Plummer, A.R., Sahinkaya, M.N., 2009. A review of wave energy converter technology. Proc. Inst. Mech. Eng. A 223 (8), 887–902. http://dx.doi.org/10.1243/09576509JPE782.

Faedo, N., Olaya, S., Ringwood, J.V., 2017. Optimal control, MPC and MPC-like algorithms for wave energy systems: An overview. IFAC J. Syst. Control 1, 37–56.

Falnes, J., Kurniawan, A., 2020. Ocean Waves and Oscillating Systems: Linear Interactions Including Wave-Energy Extraction, vol. 8, Cambridge University Press.

Fusco, F., Ringwood, J.V., 2010. Short-term wave forecasting for real-time control of wave energy converters. IEEE Trans. Sustain. Energy 1 (2), 99–106.

Fusco, F., Ringwood, J.V., 2012. A simple and effective real-time controller for wave energy converters. IEEE Trans. Sustain. Energy 4 (1), 21–30.

Gaspar, J.F., Kamarlouei, M., Sinha, A., Xu, H., Calvário, M., Faÿ, F.-X., Robles, E., Soares, C.G., 2016. Speed control of oil-hydraulic power take-off system for oscillating body type wave energy converters. Renew. Energy 97, 769–783.

Genest, R., Ringwood, J.V., 2016. A critical comparison of model-predictive and pseudospectral control for wave energy devices. J. Ocean Eng. Mar. Energy 2, 485–499.

Hals, J., Falnes, J., Moan, T., 2010. Constrained optimal control of a heaving buoy wave-energy converter. J. Offshore Mech. Arct. Eng. 133 (1), 011401. http://dx.doi.org/10.1115/1.4001431.

Kempener, R., Neumann, F., 2014. Wave Energy Technology Brief. International Renewable Energy Agency (IRENA).

Kober, J., Bagnell, J.A., Peters, J., 2013. Reinforcement learning in robotics: A survey. Int. J. Robotics Res. 32 (11), 1238–1274.

Korde, U.A., Ringwood, J., 2016. Hydrodynamic Control of Wave Energy Devices. Cambridge University Press.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25.

Lee, C., 1995. WAMIT Theory Manual, Department of Ocean Engineering. MIT, MA.

Li, Y., 2017. Deep reinforcement learning: An overview. arXiv preprint arXiv:1701.07274.

Li, G., Belmont, M.R., 2014. Model predictive control of sea wave energy converters – Part I: A convex approach for the case of a single device. Renew. Energy 69, 453–463. http://dx.doi.org/10.1016/j.renene.2014.03.070.

Li, L., Gao, Z., Yuan, Z.-M., 2019. On the sensitivity and uncertainty of wave energy conversion with an artificial neural-network-based controller. Ocean Eng. 183, 282–293.

Li, L., Yuan, Z., Gao, Y., 2018. Maximization of energy absorption for a wave energy converter using the deep machine learning. Energy 165, 340–349.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M., 2013. Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602.

Ringwood, J.V., Bacelli, G., Fusco, F., 2014. Energy-maximizing control of wave-energy converters: The development of control system technology to optimize their operation. IEEE Control Syst. Mag. 34 (5), 30–55.

Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al., 2016. Mastering the game of Go with deep neural networks and tree search. Nature 529 (7587), 484–489.

Sutton, R.S., Barto, A.G., 1998. Introduction to Reinforcement Learning, vol. 135, MIT press Cambridge.

Sutton, R.S., Barto, A.G., 2018. Reinforcement Learning: An Introduction. MIT Press.

Tri, N.M., Truong, D.Q., Binh, P.C., Dung, D.T., Lee, S., Park, H.G., Ahn, K.K., et al., 2016. A novel control method to maximize the energy-harvesting capability of an adjustable slope angle wave energy converter. Renew. Energy 97, 518–531.

Watkins, C.J., Dayan, P., 1992. Q-learning. Mach. Learn. 8, 279–292.

Weiss, G., Li, G., Mueller, M., Townley, S., Belmont, M., 2012. Optimal control of wave energy converters using deterministic sea wave prediction. In: Fuelling the Future: Advances in Science and Technologies for Energy Generation, Transmission and Storage, vol. 396, Brown-Walker Press.

Wu, F., Jing, R., Zhang, X.-P., Wang, F., Bao, Y., 2021. A combined method of improved grey BP neural network and MEEMD-ARIMA for day-ahead wave energy forecast. IEEE Trans. Sustain. Energy 12 (4), 2404–2412.

Yu, Z., Falnes, J., 1995. State-space modelling of a vertical cylinder in heave. Appl. Ocean Res. 17 (5), 265–275.

Zhan, S., Li, G., 2019. Linear optimal noncausal control of wave energy converters. IEEE Trans. Control Syst. Technol. 27 (4), 1526–1536. http://dx.doi.org/10.1109/TCST.2018.2812740.

Zhang, X., Li, Y., Gao, S., Ren, P., 2021. Ocean wave height series prediction with numerical long short-term memory. J. Mar. Sci. Eng. 9 (5), 514.

Zhang, Y., Zeng, T., Li, G., 2019. Robust excitation force estimation and prediction for wave energy converter M4 based on adaptive sliding-mode observer. IEEE Trans. Ind. Inform. 16 (2), 1163–1171.

Zhong, Q., Yeung, R.W., 2018. An efficient convex formulation for model-predictive control on wave-energy converters. J. Offshore Mech. Arct. Eng. 140 (3), 031901.

Zou, S., Zhou, X., Khan, I., Weaver, W.W., Rahman, S., 2022. Optimization of the electricity generation of a wave energy converter using deep reinforcement learning. Ocean Eng. 244, 110363.